

# The Matsu Wheel:

0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, ...

A Cloud-based  
Scanning Framework  
for Analyzing Large Volumes of  
Hyperspectral Data

**Maria Patterson, PhD**

*Open Science Data Cloud*

*Center for Data Intensive Science (CDIS)*

*University of Chicago*

HysplRI Symposium, 5 June, 2014

The Open Science Data Cloud (OSDC) is an **open-source**, **cloud-based** infrastructure that allows scientists to manage, share, and analyze medium to large size scientific datasets.



### Total OSDC Resource Size

TOTAL COMPUTE CORES  
**7550**

COMPUTE RAM  
**27622 (GB)**

RAW STORAGE  
**10.03 (PB)**

USEABLE STORAGE  
**5.92 (PB)**

### Public Data Commons

The OSDC hosts a local mirror of **1 PB** of publically available datasets.  
The data can also be freely downloaded using rsync or UDR.

#### EXAMPLE AVAILABLE DATASETS



1000 GENOMES



MODENCODE



E01



MODIS



NCBI DATASETS



COMPLETE  
GENOMICS



US CENSUS

Application for resources available to anyone doing scientific research:

[www.opensciencedatacloud.org](http://www.opensciencedatacloud.org)

# User view: 1) login

 Console Apply Public Data Systems Projects Status Support News PIRE

 OPEN SCIENCE DATA CLOUD

## Cloud Services for the Scientific Community

The OSDC provides petabyte-scale cloud resources that let you easily analyze, manage, and share data.

[Get Started Now](#)[OSDC Console Login](#)

### Featured on the OSDC

## Project Matsu



### How can I get involved?

#### Apply

Fill out a short application for an OSDC resource allocation. Allocations start at 16 dedicated cores and 1TB of storage, but scale depending on the project needs and level of organizational partnership.

#### Partner

Partner with us and add your own racks to the OSDC (we will manage them for you). Organizations can also join the [Open Cloud Consortium \(OCC\)](#) which is made up of working groups, including the OSDC.

#### Develop

All of the software developed as part of the OSDC is open source and hosted on GitHub. You can directly help the scientific cloud computing community by contributing to the open source OSDC software stack.



# User view: 2) launch virtual machine

**OSDC** **Console** **Apply** **Public Data** **Systems** **Projects** **Status** **Support** **News** **PIRE**

Overview Instances **Images & Snapshots** Access & Security

Logged in as: mpatterson [Settings](#) [Sign Out](#)

## Images

<input type="checkbox"/>	Image Name	Type	Status	Public	Format	Cloud	Action
<input type="checkbox"/>	CentOS-5.10-v1.0	Image	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	sullivan_useme5	Image	Active	No	QCOW2	Sullivan	Launch
<input type="checkbox"/>	CentOS-6.0-v1.0	Image	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	Ubuntu-12.04-LTS-v1.3	Image	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	Ubuntu-12.04-LTS-v1.6	Image	Active	Yes	QCOW2	Atwood	Launch
<input type="checkbox"/>	Ubuntu-12.04-LTS-v1.5_ia32-libs	Image	Active	Yes	QCOW2	Atwood	Launch
<input type="checkbox"/>	Ubuntu-12.04-LTS-v1.5	Image	Active	Yes	QCOW2	Atwood	Launch

Displaying 7 items

## User Snapshots

<input type="checkbox"/>	Image Name	Type	Status	Public	Format	Cloud	Action
<input type="checkbox"/>	sullivan_nc_hdf_4	Snapshot	Active	No	QCOW2	Sullivan	Launch
<input type="checkbox"/>	sullivan_netcdf_hdf5_2	Snapshot	Active	No	QCOW2	Sullivan	Launch
<input type="checkbox"/>	sullivan_geotiffs3	Snapshot	Active	No	QCOW2	Sullivan	Launch
<input type="checkbox"/>	sullivan_useme5	Image	Active	No	QCOW2	Sullivan	Launch

## Launch Instance

Details Access & Security Post-Creation

### Instance Source

Snapshot

Specify the details for launching an instance.

### Instance Snapshot

sullivan\_netcdf\_hdf5\_2

The chart below shows the resources used by this project in relation to the project's quotas.

### Flavor Details

Name	m1.xlarge
VCPUs	8
Root Disk	20 GB
Ephemeral Disk	0 GB
Total Disk	20 GB
RAM	16,384 MB

### Instance Name

sullivan\_xlarge

### Flavor

m1.xlarge

### Instance Count

1

### Cloud Name

Sullivan

### Project Quotas

Number of Instances (0)	10 Available
Number of VCPUs (0)	16 Available
Total RAM (0 MB)	66,536 MB Available

Cancel Launch

# User view: 3) run analysis

**OSDC** **Console** **Apply** **Public Data** **Systems** **Projects** **Status** **Support** **News** **PIRE**

Overview **Instances** Images & Snapshots Access & Security

Logged in as: mpatterson [Settings](#) [Sign Out](#)

## Instances

[Launch Instance](#)[Terminate Instances](#)

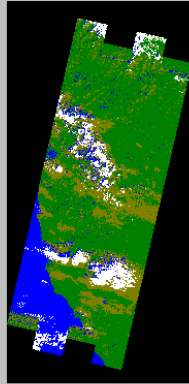
<input type="checkbox"/>	Instance Name	IP Address	Size	Keypair	Status	Task	Power State	Cloud	Actions
<input type="checkbox"/>	EO-1 analysis	172.16.1.102	m1.small   2GB RAM   1 VCPU   20GB Disk	osdc_keypair	Active	None	Running	Sullivan	<a href="#">Create Snapshot</a> ▼

Displaying 1 item

1. mpatterson@kg14-compute-1: ~ (ssh)

mpatterson@kg14-compute-1: ~  
recorded and subject to audit. Use of this system is expressed consent to such monitoring and recording. Any unauthorized access or use of this Automated Information System is prohibited and could be subject to criminal and civil penalties.  
Users are prohibited from using the system in a non approved manner. All data is to be treated as privileged information and transferred in an approved secure manner.  
mpatterson@kg14-compute-1:~\$

ubuntu@eo-1-analysis: ~/eo1\_demo  
System information as of Wed Jun 4 20:38:54 CDT 2014  
  
System load: 0.08 Processes: 65  
Usage of /: 33.3% of 19.69GB Users logged in: 0  
Memory usage: 2% IP address for eth0: 172.16.1.102  
Swap usage: 0%  
  
Graph this data and manage this system at <https://landscape.canonical.com/>  
Get cloud support with Ubuntu Advantage Cloud Guest  
<http://www.ubuntu.com/business/services/cloud>  
ubuntu@eo-1-analysis:~\$ ls  
eo1\_demo  
ubuntu@eo-1-analysis:~\$ cd eo1\_demo/  
ubuntu@eo-1-analysis:~/eo1\_demo\$ ls  
classify.py       italyClassified.tif   trainingSpectra.png  
FourClassTrainingSet.txt   makeRGB.py       viewClassifiedTiff.py  
italyClassified.png       README.md       viewGeoTiff.py  
ubuntu@eo-1-analysis:~/eo1\_demo\$ less makeRGB.py  
ubuntu@eo-1-analysis:~/eo1\_demo\$ python view  
viewClassifiedTiff.py viewGeoTiff.py  
ubuntu@eo-1-analysis:~/eo1\_demo\$ python viewClassifiedTiff.py  
usage: viewClassifiedTiff.py [-h] tifname [outfile]  
viewClassifiedTiff.py: error: too few arguments  
ubuntu@eo-1-analysis:~/eo1\_demo\$ python viewClassifiedTiff.py italyClassified.tif

Figure 1  
  
zoom reset

**Compute for general projects +  
selected groups**

**"Adler"**

(retiring)

38 nodes

288 cores

1134 GB RAM

312 TB

**"Sullivan"**

58 nodes

1856 cores

4736 GB RAM

435 TB

**"Goldberg"**

60 nodes

1056 cores

4424 GB RAM

544 TB

**Protected compute  
clouds**

**"Bionimbus-  
PDC"**

100 nodes

2336 cores

9856 GB RAM

1313 TB

**"Atwood"**

23 nodes

184 cores

736 GB RAM

196 TB



**Public Data  
Commons**

~ 1 PB storage  
(no compute)

**Hadoop clusters  
for selected projects**

**"Skidmore"  
(Matsu Wheel)**

25 nodes

800 cores

3200 GB RAM

261 TB

**"OCC-Y"**

61 nodes

976 cores

3184 GB RAM

1101 TB

**"LVOC"**

(Matsu, retiring)

9 nodes

54 cores

352 GB RAM

72 TB

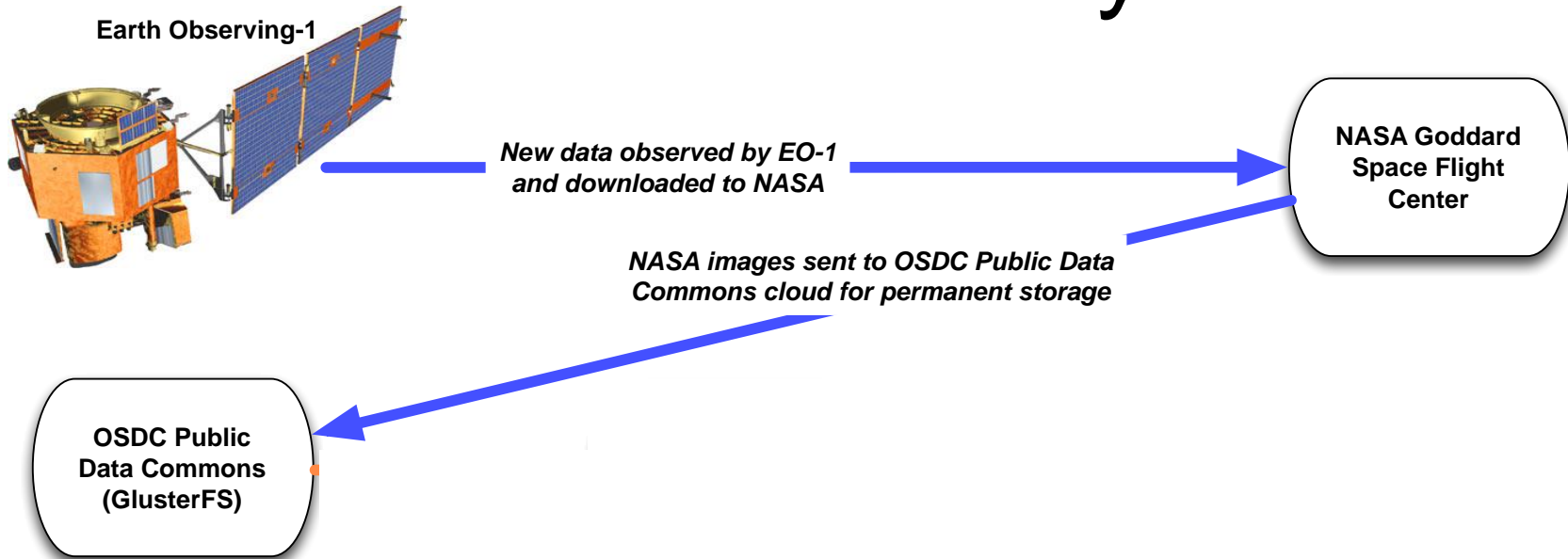


# Project Matsu



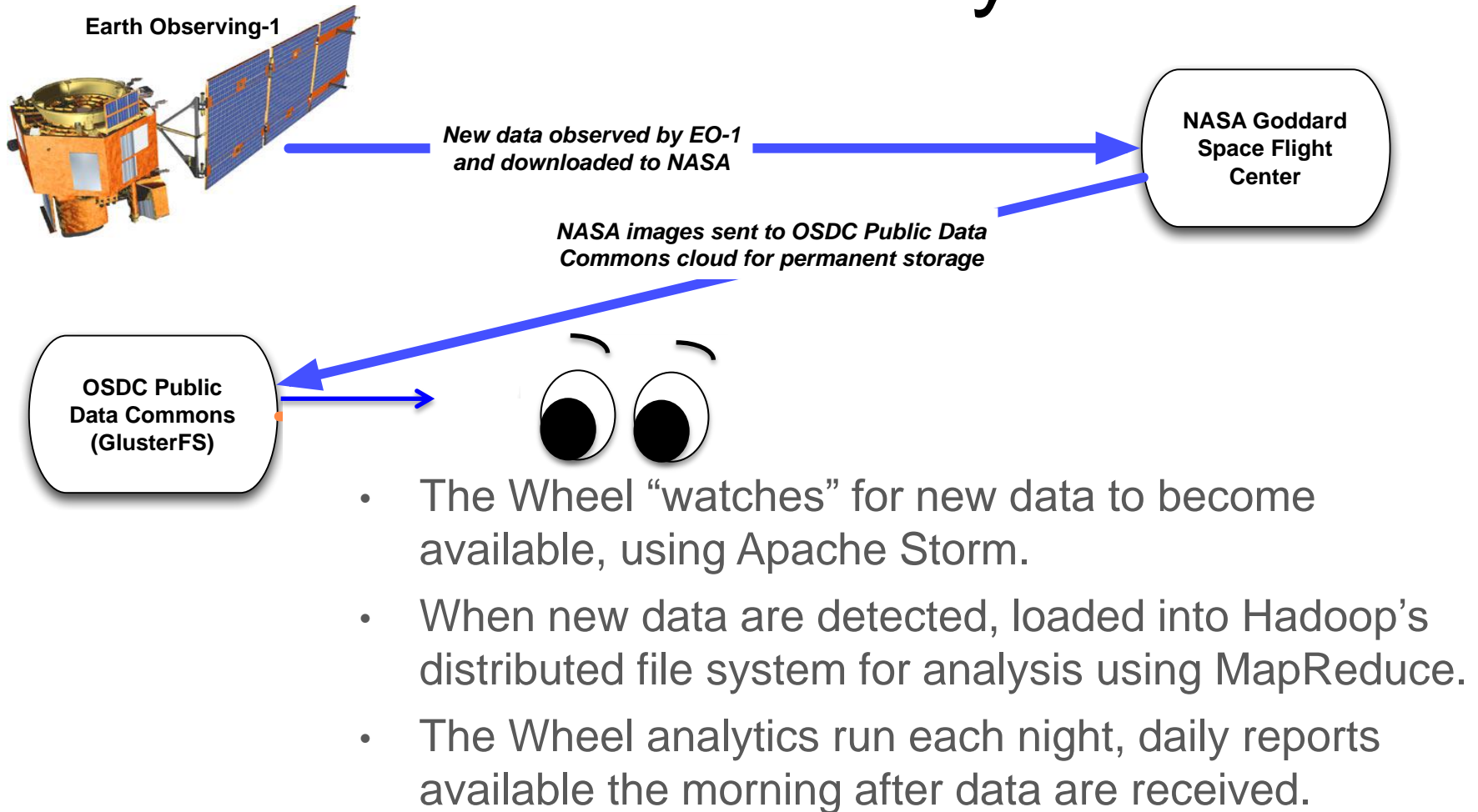
- Joint effort between the Open Cloud Consortium (lead, Robert Grossman) and NASA (lead, Dan Mandl) to develop open source technology for cloud-based processing of satellite imagery to support earth sciences.
- The OSDC is used to process Earth Observing 1 (EO-1) *satellite imagery* from the Advanced Land Imager and the Hyperion instruments and to make this data available to interested users.
  - Namibia flood dashboard, WCPS
  - Hadoop-based 'Matsu Wheel' scanning data algorithm

# Matsu Analytic Wheel

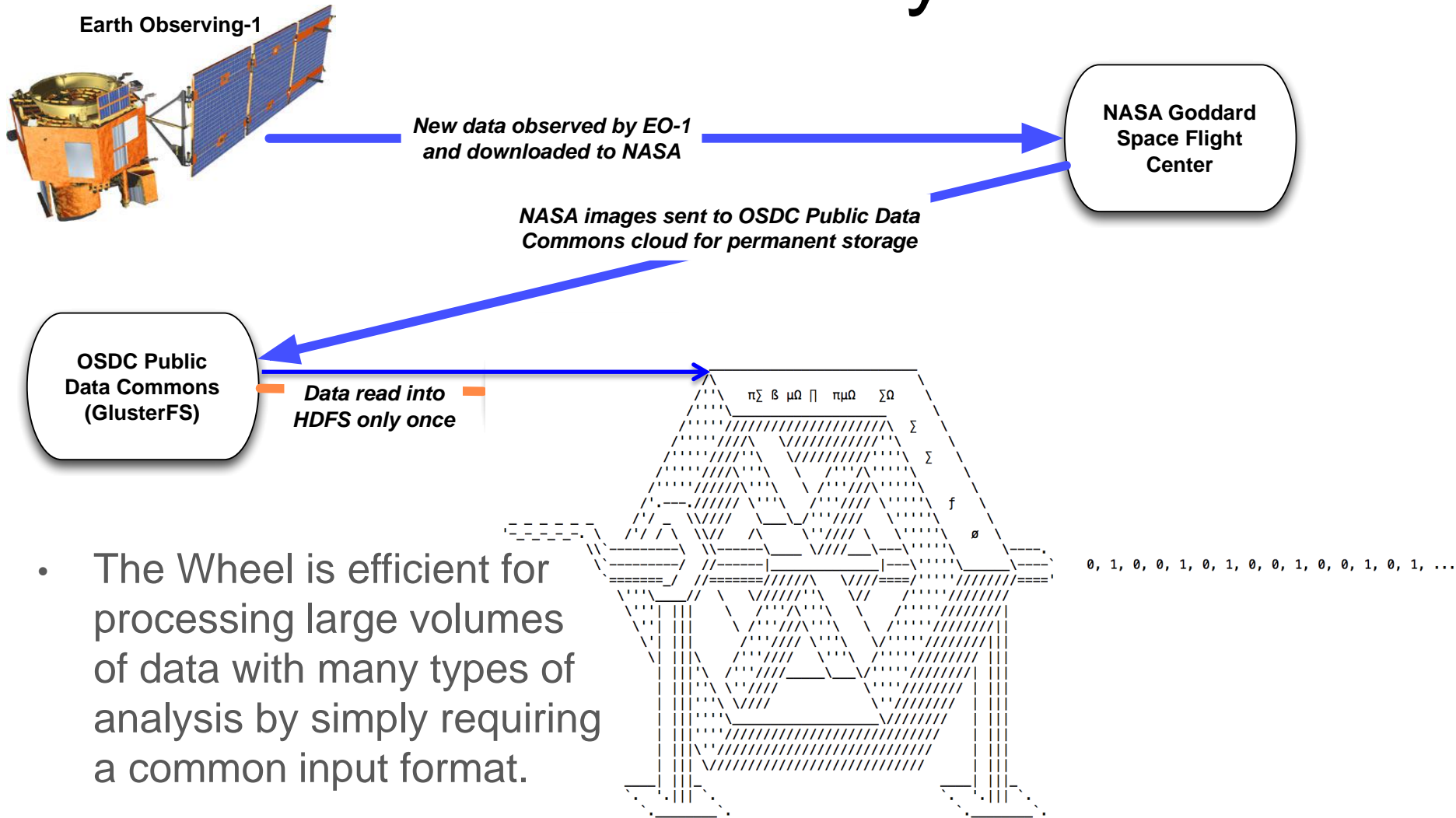




# Matsu Analytic Wheel

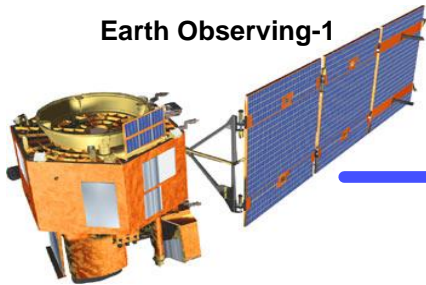


# Matsu Analytic Wheel



- The Wheel is efficient for processing large volumes of data with many types of analysis by simply requiring a common input format.

# Matsu Analytic Wheel



Earth Observing-1

*New data observed by EO-1  
and downloaded to NASA*

**NASA Goddard  
Space Flight  
Center**

*NASA images sent to OSDC Public Data  
Commons cloud for permanent storage*

**OSDC Public  
Data Commons  
(GlusterFS)**

*Data read into  
HDFS only once*

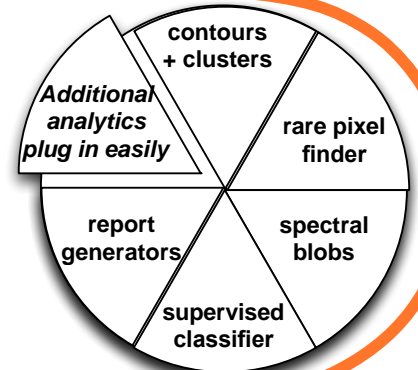
**HDFS**

*Metadata stored*

**NoSql Database  
(Accumulo)**

*Analytic results stored*

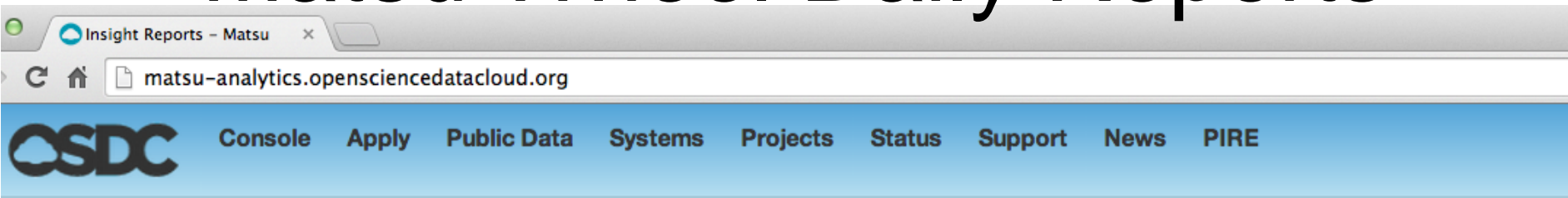
*Wheel analytics run  
over data using MapReduce*



*Analytic reports generated by  
Wheel are accessible via web browser*

*Secondary  
analysis can be  
done from  
analytic database*

# Matsu Wheel Daily Reports



EO-1 Level 0 images are received daily from NASA and are transformed into various Level 1 products. This transformation does not run each day. Level 1G images are converted (SequenceFiles), uploaded (HDFS), and MapReduced (analytic) once a day to build the Insight Reports. If the Level 1G is not available at that time, it is processed the next day, but not all days have images. Daily processing began in July 2013. Previous to that, a few select days from 2010 and 2012 were processed.

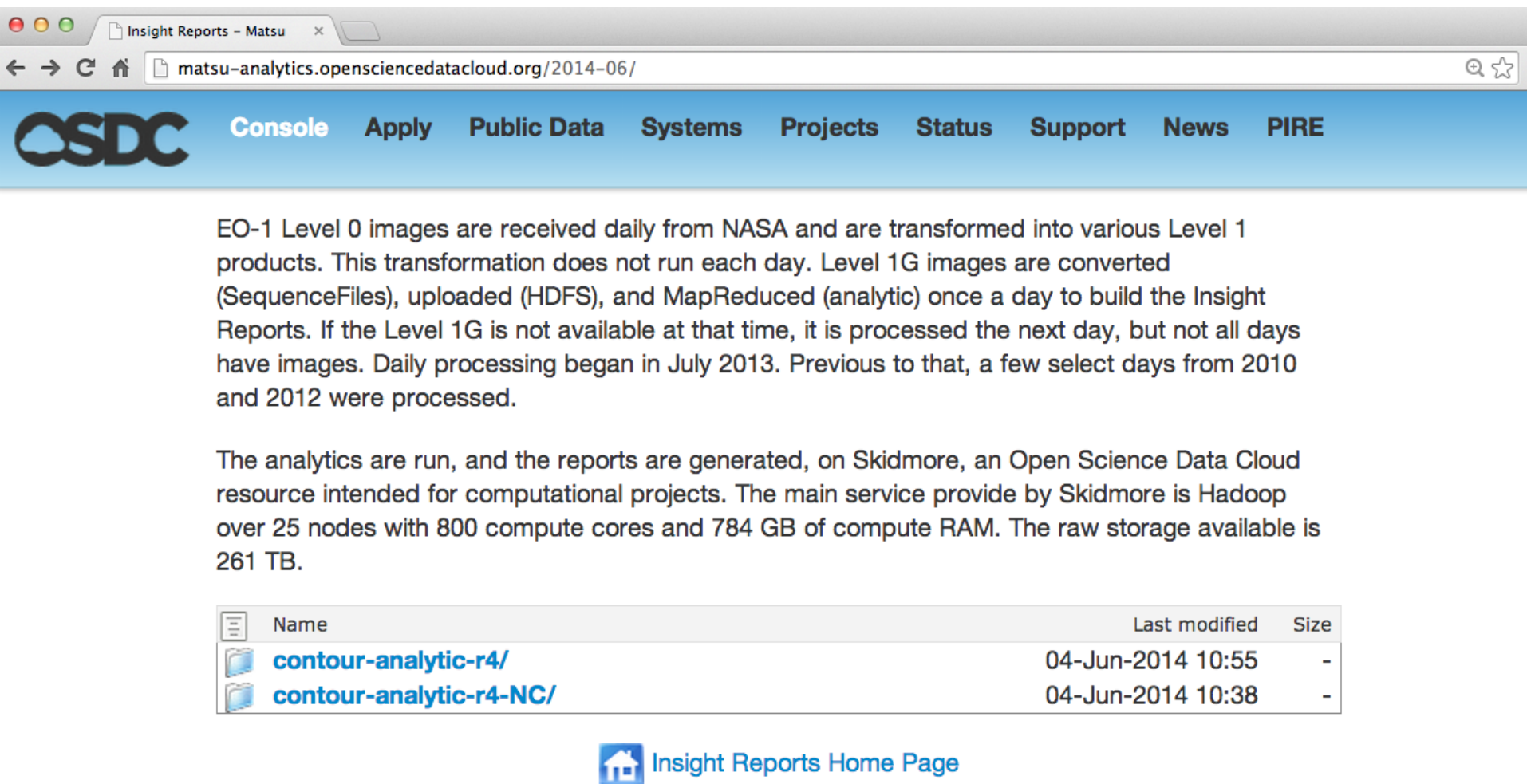
The analytics are run, and the reports are generated, on Skidmore, an Open Science Data Cloud resource intended for computational projects. The main service provide by Skidmore is Hadoop over 25 nodes with 800 compute cores and 784 GB of compute RAM. The raw storage available is 261 TB.

Name	Last modified	Size
2014-06/	04-Jun-2014 10:18	-
2014-05/	01-Jun-2014 10:23	-
2014-04/	20-Apr-2014 12:23	-
2014-03/	20-Apr-2014 11:47	-
2014-02/	20-Apr-2014 11:06	-
2014-01/	16-Apr-2014 16:13	-
2013-12/	25-Mar-2014 17:18	-
2013-11/	25-Mar-2014 17:18	-
2013-10/	25-Mar-2014 17:18	-
2013-09/	25-Mar-2014 17:18	-
2013-08/	25-Mar-2014 17:18	-
2013-07/	25-Mar-2014 17:18	-
2012/	25-Mar-2014 17:18	-
2010/	25-Mar-2014 17:18	-

matsu-analytics.opensciencedatacloud.org




# Matsu Wheel Daily Reports



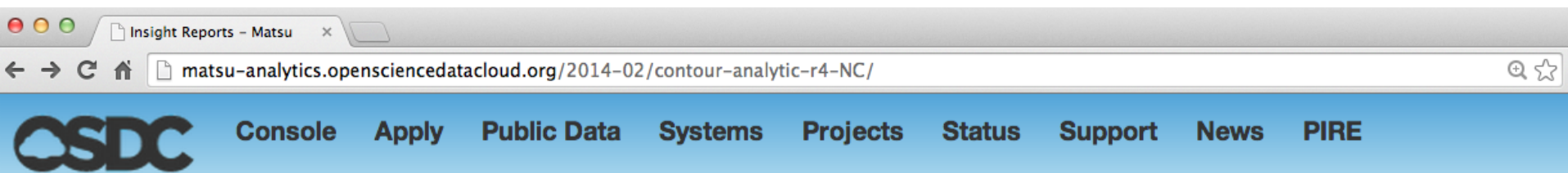
EO-1 Level 0 images are received daily from NASA and are transformed into various Level 1 products. This transformation does not run each day. Level 1G images are converted (SequenceFiles), uploaded (HDFS), and MapReduced (analytic) once a day to build the Insight Reports. If the Level 1G is not available at that time, it is processed the next day, but not all days have images. Daily processing began in July 2013. Previous to that, a few select days from 2010 and 2012 were processed.

The analytics are run, and the reports are generated, on Skidmore, an Open Science Data Cloud resource intended for computational projects. The main service provide by Skidmore is Hadoop over 25 nodes with 800 compute cores and 784 GB of compute RAM. The raw storage available is 261 TB.

Name	Last modified	Size
<a href="#">contour-analytic-r4/</a>	04-Jun-2014 10:55	-
<a href="#">contour-analytic-r4-NC/</a>	04-Jun-2014 10:38	-







 [Insight Reports Home Page](#)

# Matsu Wheel Daily Reports

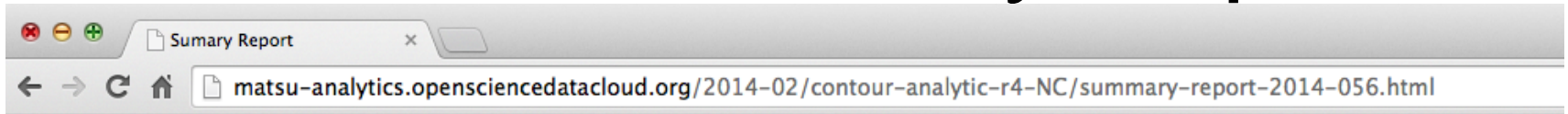


EO-1 Level 0 images are received daily from NASA and are transformed into various Level 1 products. This transformation does not run each day. Level 1G images are converted (SequenceFiles), uploaded (HDFS), and MapReduced (analytic) once a day to build the Insight Reports. If the Level 1G is not available at that time, it is processed the next day, but not all days have images. Daily processing began in July 2013. Previous to that, a few select days from 2010 and 2012 were processed.

The analytics are run, and the reports are generated, on Skidmore, an Open Science Data Cloud resource intended for computational projects. The main service provide by Skidmore is Hadoop over 25 nodes with 800 compute cores and 784 GB of compute RAM. The raw storage available is 261 TB.

Name	Last modified	Size
 <a href="#">overlays/</a>	17-Apr-2014 12:34	-
 <a href="#">summary-report-2014-059.html</a>	17-Apr-2014 03:12	7.3K
 <a href="#">summary-report-2014-058.html</a>	17-Apr-2014 02:43	7.5K
 <a href="#">summary-report-2014-057.html</a>	17-Apr-2014 02:14	7.7K
 <a href="#">summary-report-2014-056.html</a>	17-Apr-2014 01:45	8.1K
 <a href="#">summary-report-2014-054.html</a>	17-Apr-2014 01:10	7.7K

# Matsu Wheel Daily Reports



## Analytic Summary Report

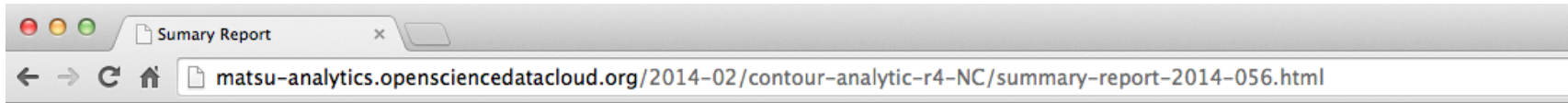
Collection Date	2014-02-25 (day 056)
Analysis Date	Thu Apr 17 01:45:32 2014
Analytic Environment	
Analytic	Contours-2013-12-r4
Noise Correction Enabled	True
Summary Stats	ss-2013-12-r1
Data Ingest	populateHDFS-2013-11-r1
Report Format	reportContoursR4
Run Summary	
Number of Image	16
Average Number of Pixels	922639
Image with Largest Variance	E01H0260472014056110PF_HYP_L1G

# Matsu Wheel Daily Reports

Summary Report
matsu-analytics.opensciencedatacloud.org/2014-02/contour-analytic-r4-NC/summary-report-2014-056.html
Images
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0110282014056110K3_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0110282014056110K3_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0250282014056110K2_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0250282014056110K2_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0260472014056110PF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0260472014056110PF_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0390092014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0390092014056110KF_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0750222014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0750222014056110KF_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0750912014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H0750912014056110KF_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1130082014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1130082014056110KF_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1160302014056110P0_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1160302014056110P0_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1281082014056110K0_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1281082014056110K0_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1491082014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1491082014056110KF_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1660262014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1660262014056110KF_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1800742014056110K3_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1800742014056110K3_HYP_L1G</a>
<a href="/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1932182014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eol/hyperion_11g/2014/056/EO1H1932182014056110KF_HYP_L1G</a>

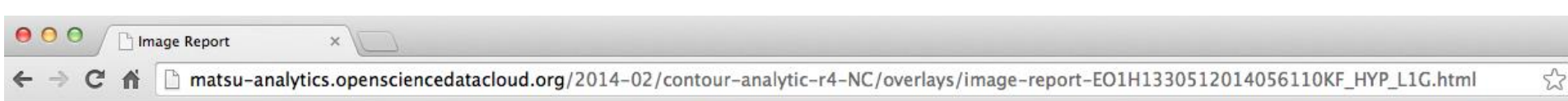


# Matsu Wheel Daily Reports



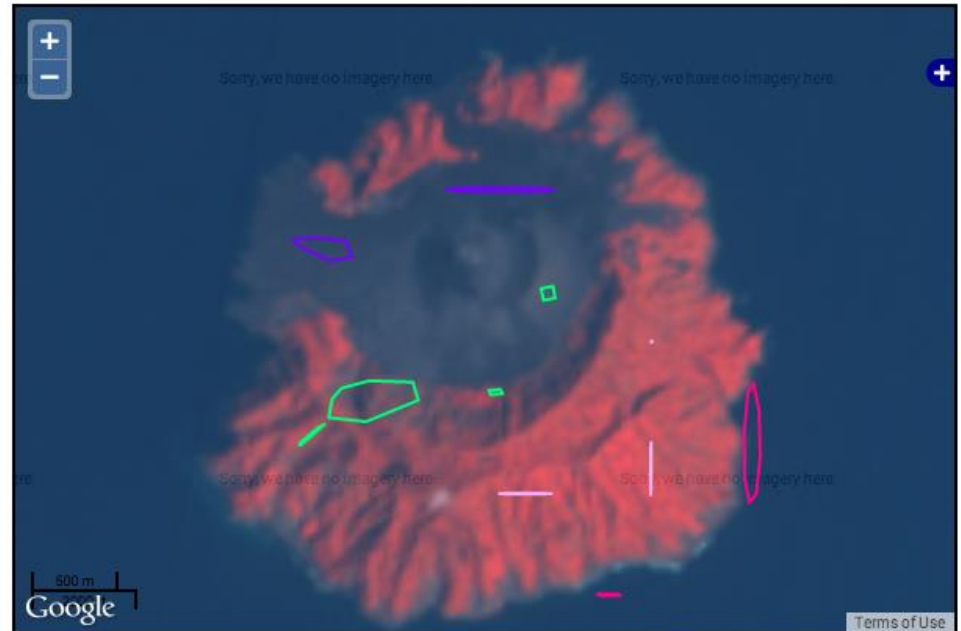
Hyper-spectral Objects					
Rank	Object Name	Cluster Score	Contour Score	Location (lng/lat)	Image
1	C0-33051-OKF	1000	0.7446	93.8739019579,12.2704698125	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
2	C2-33051-OKF	1000	0.6196	93.8682455446,12.2799335922	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
3	C2-33051-OKF	1000	0.4510	93.8585620292,12.2742192321	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
4	C2-33051-OKF	1000	0.4176	93.8552204394,12.2723263195	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
5	C0-33051-OKF	1000	0.3845	93.8670069173,12.2691974168	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
6	C3-33051-OKF	1000	0.3512	93.8656741679,12.2854892912	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
7	C0-33051-OKF	1000	0.2894	93.8752966894,12.2775235269	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
8	C2-33051-OKF	1000	0.2361	93.8653887507,12.2746401012	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H1330512014056110KF_HYP_L1G</a>
9	C1-26086-OK1	1000	0.1580	-61.3432105294,-37.5568381617	<a href="/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H2260862014056110K1_HYP_L1G">/glusterfs/osdc_public_data/eo1/hyperion_l1g/2014/056/EO1H2260862014056110K1_HYP_L1G</a>
	C2-				

# Matsu Wheel Daily Reports



## Matsu Analytic Image Report

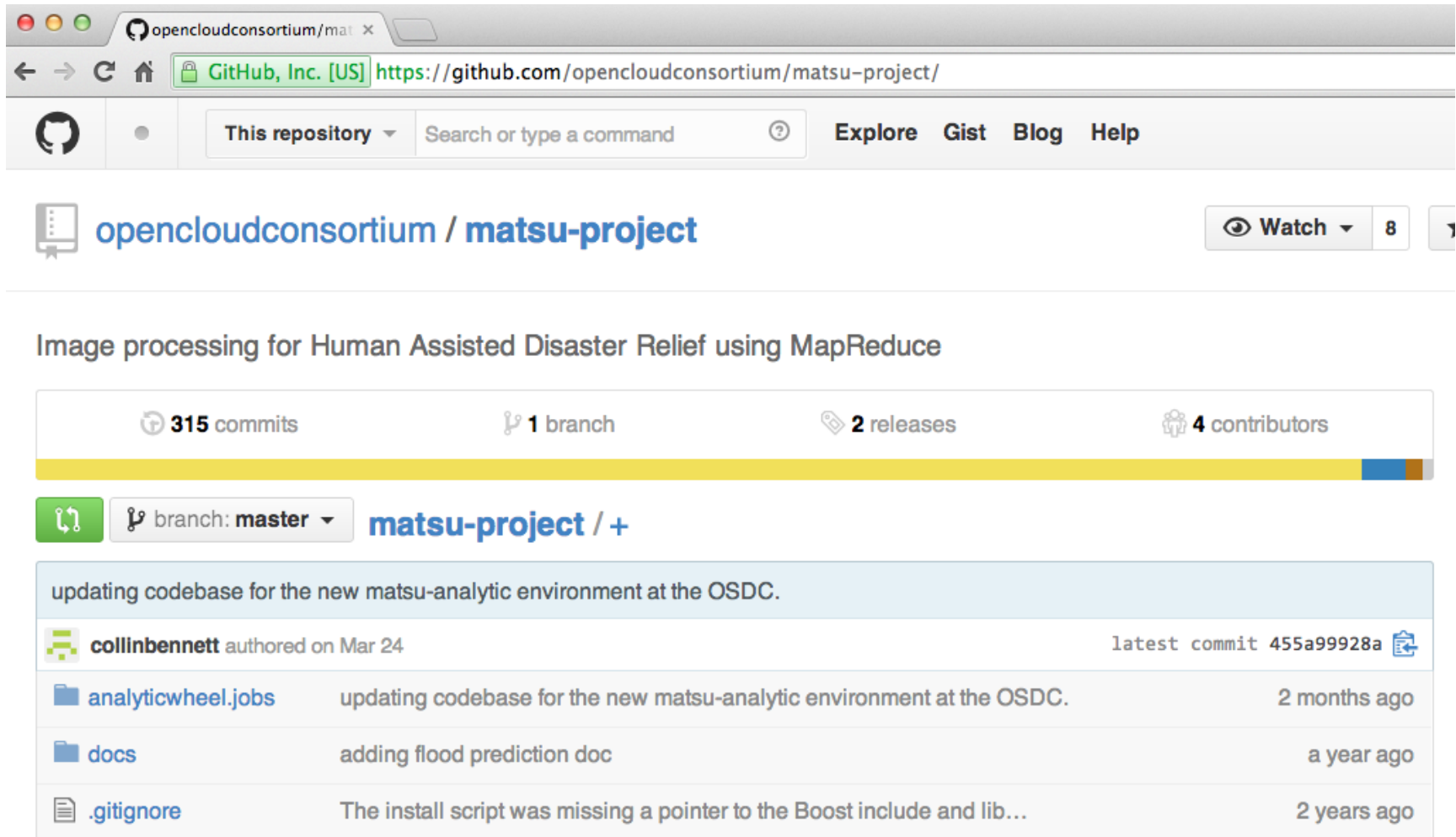
Collection Date	2014-02-25 (day 056)
Analysis Date	Thu Apr 17 01:45:31 2014
Analytic Environment	
Analytic	Contours-2013-12-r4
Noise Correction Enabled	True
Summary Stats	ss-2013-12-r1
Data Ingest	populateHDFS-2013-11-r1
Report Format	reportContoursR4
Hyperspectral Image	
Image	EO1H1330512014056110KF_HYP_L1G
Number of Bands	242



Contour ID	Cluster Score	Contour Score	lat,lng	Area (Pixels)	Area (Meters)	color	Spectral Signature
C2-33051-0KF	1000	0.4176	93.8552204394,12.2723263195	2.1740	137.7266	COLOR	<a href="#">wavelengths</a>
C2-33051-0KF	1000	0.4510	93.8585620292,12.2742192321	103.3753	6549.0126	COLOR	<a href="#">wavelengths</a>



# Matsu Wheel is open source



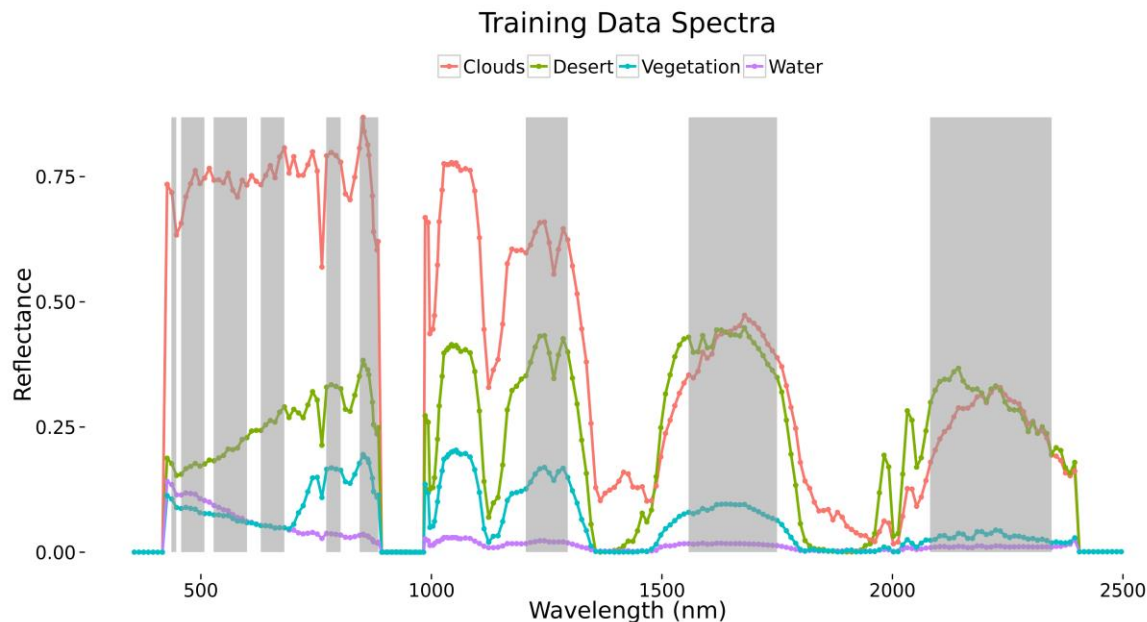
The screenshot shows the GitHub repository page for `opencloudconsortium/matsu-project`. The repository is titled "Image processing for Human Assisted Disaster Relief using MapReduce". It has 315 commits, 1 branch, 2 releases, and 4 contributors. The current branch is `master`. The repository description is "updating codebase for the new matsu-analytic environment at the OSDC." The latest commit is by `collinbennett` on Mar 24, with commit hash `455a99928a`. The repository contains the following files and folders:

File/Folder	Description	Last Commit
<code>analyticwheel.jobs</code>	updating codebase for the new matsu-analytic environment at the OSDC.	2 months ago
<code>docs</code>	adding flood prediction doc	a year ago
<code>.gitignore</code>	The install script was missing a pointer to the Boost include and lib...	2 years ago

[github.com/opencloudconsortium/matsu-project](https://github.com/opencloudconsortium/matsu-project)

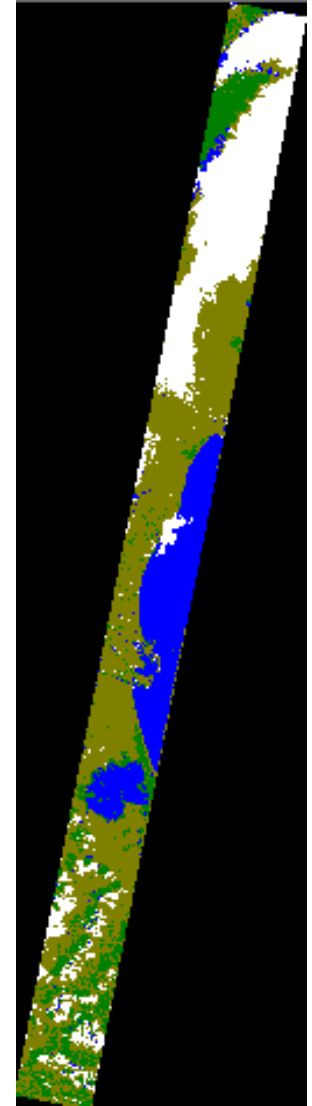
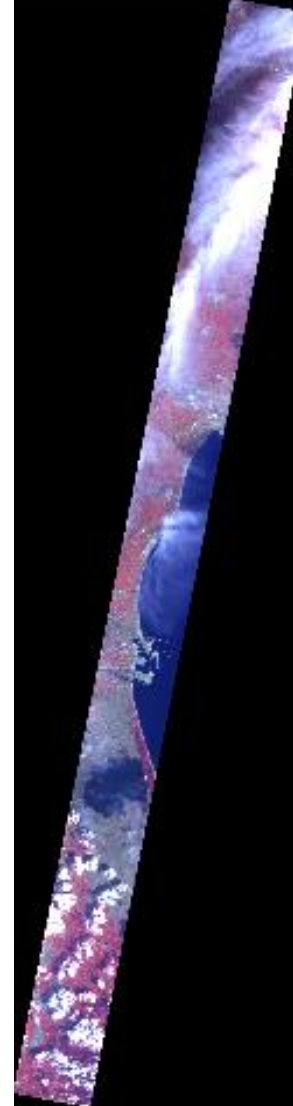
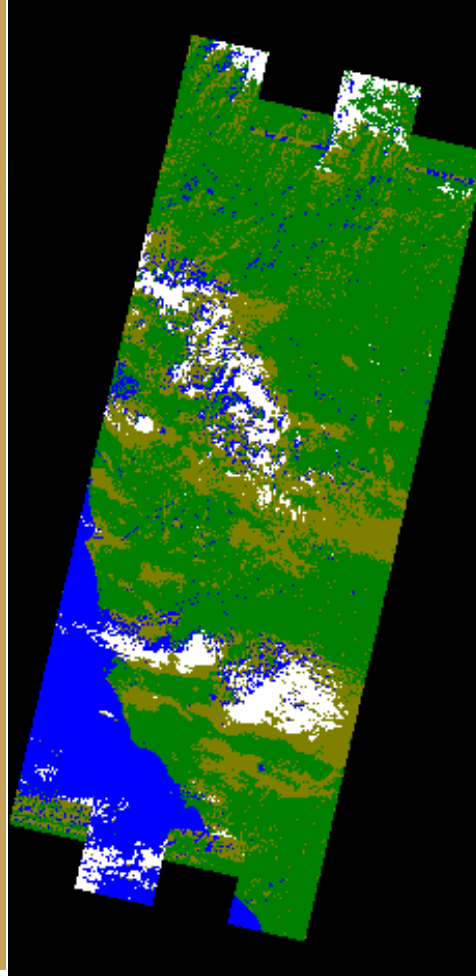
# New wheel analytic (beta): Support Vector Machine (SVM) classifier

- A supervised machine learning classification algorithm
- Train the classifier by hand classifying areas in a set of training images
- Beta classifier has 4 classes: clouds, dry land, vegetation, water





# New wheel analytic (beta): Support Vector Machine (SVM) classifier



# Continuing work

- SVM classifier adapt regionally to geographic area (classes depend on geography)
- Incorporate SVM classifier into Matsu Wheel
- Additional wheel analytics
- Web Map Service and tiling using Geoserver
- Add additional data to the Wheel

## What you can do

- Make your data available to Project Matsu
- Port your analysis tools and applications
- Use the Matsu cloud to facilitate making discoveries that require integrating multiple large datasets
- Contribute a Wheel analytic